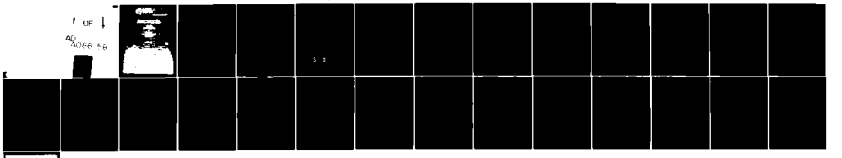
AD-A086 158 NAVAL RESEARCH LAB WASHINGTON DC F/G 9/2
MINIMUM CROSS-ENTROPY PATTERN CLASSIFICATION AND CLUSTER ANALYS--ETC(U)
APR 80 J E SHORE, R M GRAY
UNCLASSIFIED NRL-MR-4207 SBIE-AD-E000 434 NL



END
DATE FILMED
8-80
DTIC

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER NRL Memorandum Report 4207	2 GOVT ACCESSION NO. AD-A086158	3 RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle) MINIMUM CROSS-ENTROPY PATTERN CLASSIFICATION AND CLUSTER ANALYSIS	5 TYPE OF REPORT & PERIOD COVERED Interim report on a continuing NRL problem. 6 PERFORMING ORG. REPORT NUMBER
7 AUTHOR(s) John E. Shore and Robert M. Gray*	8 CONTRACT OR GRANT NUMBER(s)
9 PERFORMING ORGANIZATION NAME AND ADDRESS Naval Research Laboratory Washington, D.C. 20375	10 PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N; RR014-09-41; ~~75~~-0102-0-0
11 CONTROLLING OFFICE NAME AND ADDRESS Naval Research Laboratory Washington, D.C. 20375	12 REPORT DATE April 10, 1980 13 NUMBER OF PAGES 25
14 MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15 SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16 DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17 DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18 SUPPLEMENTARY NOTES

*Present address: Department of Electrical Engineering, Stanford University, Stanford, California.

This research was partially supported by the Office of Naval Research and by the National Science Foundation.

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Cross-entropy
Discrimination information
Pattern recognition
Cluster analysis
Information theory

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper considers the problem of classifying an input vector of measurements by a nearest-neighbor rule applied to a fixed set of vectors. The fixed vectors are sometimes called characteristic feature vectors, codewords, cluster centers, models, reproductions, etc. The nearest-neighbor rule considered uses a non-Euclidean, information-theoretic distortion measure that is not a metric, but that nevertheless leads to a classification method that is optimal in a well-defined sense and is also computationally attractive. Furthermore, the distortion measure results in a simple method of computing

(Continues)

20. Abstract (Continued)

cluster centroids. Our approach is based on cross-entropy minimization (also called minimum discrimination information or minimum directed divergence), and can be viewed as a refinement of a general classification method due to Kullback. The refinement exploits special properties of cross-entropy that hold when the probability densities involved happen to be minimum cross-entropy densities. The approach is a generalization of a recently-developed speech coding technique.

CONTENTS

MINIMUM CROSS-ENTROPY PATTERN CLASSIFICATION AND CLUSTER ANALYSIS

I. INTRODUCTION AND STATEMENT OF THE PROBLEM

Let $\underset{\sim}{F}$ denote a "feature vector" of measurements F_i $(i = 0,1,\ldots,M)$ that are made on a system for which the individual measurements can be expressed as expected values with respect to some unknown underlying probability density function $q^\dagger(\underset{\sim}{x})$:

$$\int d\underset{\sim}{x}\ f_i(\underset{\sim}{x}) q^\dagger(\underset{\sim}{x}) \ = \ F_i \ , \tag{1}$$

where the f_i are known functions and $\underset{\sim}{x}$ is a finite dimensional vector. This paper considers the problem of classifying $\underset{\sim}{F}$ by identifying the vector $\hat{\underset{\sim}{F}}^{(t)}$ that "best represents" $\underset{\sim}{F}$ according to the "nearest-neighbor rule"

$$D(\underset{\sim}{F},\hat{\underset{\sim}{F}}^{(t)}) \ = \ \min_{s\in\Lambda} D(\underset{\sim}{F},\hat{\underset{\sim}{F}}^{(s)}) \quad , \tag{2}$$

where D is a distortion measure and where $\{\hat{\underset{\sim}{F}}^{(s)} : s\in\Lambda\}$ is a discrete or continuous set of pre-defined vectors. The vectors $\hat{\underset{\sim}{F}}^{(s)}$ might be called characteristic feature vectors, codewords, cluster centers, models, reproductions, etc. An example of such a problem occurs in speech analysis, where the measurements F_i are estimates of autocorrelation function values, which can be expressed as expectations with respect to some underlying distribution [1], [2]. In speech recognition applications, the identity of the best codeword $\hat{\underset{\sim}{F}}^{(s)}$ could be used to identify the speech sound or perhaps the speaker; in speech transmission applications, the identity of the best codeword can be transmitted as part of a narrow bandwidth encoding of the speech [2],[3].

Most of the literature on nearest-neighbor classification deals only with Euclidean or other metric distortion measures [4],[5]. In contrast, we consider an information-theoretic distortion measure that is not a metric, but that nevertheless leads to a classification algorithm that is optimal in a

Manuscript submitted February 21, 1980.

well-defined sense and is also computationally attractive. Furthermore, the distortion measure results in a simple method of computing cluster centroids. Our approach exploits properties of the cross-entropy between any two probability density functions q,p, defined by

$$H[q,p] = \int d\underset{\sim}{x}\; q(\underset{\sim}{x}) \log(q(\underset{\sim}{x})/p(\underset{\sim}{x})) \;, \tag{3}$$

if the measure induced by q is absolutely continuous with respect to that induced by p, and $H[q,p] = \infty$ otherwise [6],[7]. In particular, our approach is based on cross-entropy having unique properties as an information measure [6]-[8] and on cross-entropy minimization having unique properties as an inference procedure [9]. The approach can be viewed as a refinement of a general classification method due to Kullback [6, p. 83]. The refinement exploits special properties of cross-entropy that hold when the probability densities involved happen to be minimum cross-entropy densitites [10],[11]. The approach is a generalization of speech coding by vector quantization [2],[3].

Section II reviews relevant properties of cross-entropy and cross-entropy minimization and Section III presents the minimum cross-entropy solution to the classification problem. Section IV considers the cluster analysis problem of choosing appropriate feature vectors $\hat{\underset{\sim}{F}}^{(s)}$. An example concerning narrow-band speech transmission is disscussed in Section V, and a general discussion follows in Section VI.

II. THEORETICAL BACKGROUND

Suppose you have a <u>prior</u> estimate p of the unknown probability density $q^{\dagger}(\underset{\sim}{x})$, you obtain new information about $q^{\dagger}$ in the form of expected value <u>constraints</u> (1), and you need to choose a <u>posterior</u> estimate q that is in some sense the best estimate of $q^{\dagger}$ given what you know. Which one should you

choose? The principle of minimum cross-entropy provides a general solution to this inference problem [9]. The principle states that, of all the distributions that satisfy the constraints, you should choose the posterior q with the least cross-entropy (3) with respect to the prior p. As a general method of statistical inference, cross-entropy minimization was first introduced by Kullback [6]. The name cross-entropy is due to Good [12]. Other names include expected weight of evidence [13, p. 72], directed divergence [6, p. 7], discrimination information [6, p. 37], and the entropy of one distribution relative to another [7, p. 19]. The principle of maximum entropy [14],[15] is a special case of cross-entropy minimization under appropriate conditions [2],[9].

A. Minimum Cross-Entropy Probability Densities

Given a positive prior probability density p, if there exists a posterior that minimizes the cross-entropy (3) and satisfies the constraints (1) and

$$\int d\underset{\sim}{x}\; q(\underset{\sim}{x}) = 1 \;, \tag{4}$$

then it has the form

$$q(\underset{\sim}{x}) = p(\underset{\sim}{x})\, \exp\left(-\lambda - \sum_{k=0}^{M} \beta_k f_k(\underset{\sim}{x})\right), \tag{5}$$

with the possible exception of a set of points on which the constraints imply that q vanishes [6, p. 38],[10]. In (5), β_k and λ are Lagrangian multipliers whose values are determined by the constraints (1) and (4). Conversely, if one can find values for β_k and λ in (5) such that the constraints (1) and (4) are satisfied, then the solution exists and is given by (5) [10]. Conditions for the existence of solutions are discussed by Csiszàr [10]. The cross-entropy at the minimum can be expressed in terms of the Lagrangian

multipliers and the F_k as follows ([6, p. 38], [11]):

$$H[q,p] = -\lambda - \sum_{k=0}^{M} \beta_k F_k \quad . \tag{6}$$

It is necessary to choose λ and the β_k so that the constraints are satisfied. In the presence of the constraint (4), one may rewrite the remaining constraints (1) in the form

$$\int d\underset{\sim}{x}\ (f_i(\underset{\sim}{x}) - F_i)q(\underset{\sim}{x}) = 0 \tag{7}$$

Now, if one finds values for the β_k such that

$$\int d\underset{\sim}{x}\ (f_i(\underset{\sim}{x}) - F_i)p(\underset{\sim}{x})\ \exp\left(-\sum_{k=0}^{M}\beta_k f_k(\underset{\sim}{x})\right) = 0 \quad , (i = 0,\ldots,M), \tag{8}$$

holds, (7) will be satisfied, and (4) can then be satisfied by setting

$$\lambda = \log \int d\underset{\sim}{x}\ p(\underset{\sim}{x})\ \exp\left(-\sum_{k=0}^{M}\beta_k f_k(\underset{\sim}{x})\right) \quad . \tag{9}$$

If the integral in (9) can be performed, one can sometimes find values for the β_k from the relations

$$-\frac{\partial\lambda}{\partial\beta_k} = F_k \quad .$$

It unfortunately is usually impossible to solve this or (8) for the β_k explicitly, in order to obtain a closed-form solution expressed directly in terms of the known expected values F_k rather than in terms of the Lagrangian multipliers. Computational methods for finding approximate solutions are, however, available ([11], [16]).

B. Justification and Properties of Cross-Entropy Minimization

In this Section, we discuss justifications for the principle of minimum

cross-entropy, and we summarize three important properties of cross-entropy minimization that lead to the classification method described in Section III. For general statements and proofs of these and other properties, see [11].

In what sense does cross-entropy minimization yield the best estimate of $q^{\dagger}$? To answer this question, it is useful and convenient to view cross-entropy minimization as one implementation of an abstract information operator $\circ$ that takes two arguments --- a prior and new information --- and yields a posterior. Thus, we write the posterior q as $q = p \circ I$, where I stands for the known constraints (1) on expected values plus the usual normalization constraint (4). Recent work has shown that, if the operator $\circ$ is required to satisfy certain axioms of consistent inference, and if $\circ$ is implemented by means of functional minimization, then the principle of minimum cross-entropy follows necessarily [9]. Informally, the axioms of $\circ$ may be phrased as follows:

1) Uniqueness. The results of taking new information into account should be unique.
2) Invariance. It shouldn't matter in which coordinate system one accounts for new information.
3) System independence. It shouldn't matter whether one accounts for independent information about independent systems separately in terms of different probability densities or together in terms of a joint density.
4) Subset Independence. It shouldn't matter whether one accounts for information about an independent subset of system states in terms of a separate conditional density or in terms of the full system density.

For the formal statements, see [9]. In terms of these axioms, the principle of cross-entropy minimization is correct in the following sense: Given a

prior probability density and new information in the form of constraints on expected values, there is only one posterior density satisfying these constraints that can be chosen in a manner that satisfies the axioms; this unique posterior can be obtained by minimizing cross-entropy.

An additional interpretation of the sense in which $q = p \circ I$ is the best estimate of $q^{\dagger}$ rests on cross-entropy's well-known [6] and unique [8] properties as an information measure. For example, cross-entropy satisfies

$$H[q,p] \geq 0 , \tag{10}$$

with equality only if $p = q$ almost everywhere. Also, if the space on which p and q are defined is the product of two sample spaces X_1 and X_2, and if p and q have the product form

$$p(x_1,x_2) = p_1(x_1)p_2(x_2)$$

and

$$q(x_1,x_2) = q_1(x_1)q_2(x_2),$$

then

$$H[q,p] = H[q_1q_2,p_1p_2] = H[q_1,p_1] + H[q_2,p_2]$$

holds. Informally speaking, H[q,p] is a measure of the "information divergence" or "information disimilarity" between q and p. In these terms, one can interpret the principle of minimum cross-entropy as follows: Since $q = p \circ I$ minimizes H[q,p], the posterior hypothesis for $q^{\dagger}$ is as close as possible in an information-measure sense to the prior hypothesis while at the same time satisfying the new constraints I. Owing to cross-entropy's properties as an information measure, H[q,p] has been proposed as a measure of the distortion introduced if p is used instead of q [17], even though H does not have properties of a metric. (For example, it does not satisfy a general triangle inequality). In the context of cross-entropy minimization, however,

there is a much stronger justification for using cross-entropy as a distortion measure. In particular, the following property holds (see [10], [11]):

Property A (triangle equality). Let I be the constraints (1) and let p be any prior. Then

$$H[q^{\dagger},p] = H[q^{\dagger},p\circ I] + H[p\circ I,p] \quad . \tag{11}$$

Thus, the minimum cross-entropy posterior estimate of $q^{\dagger}$ is not only logically consistent, but also closer to $q^{\dagger}$, in the cross-entropy sense, than is the prior p. Moreover, the difference $H[q^{\dagger},p]-H[q^{\dagger},p\circ I]$ is exactly the cross-entropy $H[p\circ I,p]$ between the posterior and the prior. Hence, $H[p\circ I,p]$ can be interpreted as the amount of information provided by the constraints I that is not inherent in p. Stated differently, $H[p\circ I,p]$ is the amount of additional distortion introduced if p is used instead of $p\circ I$. Since, for any density r there exist constraints I_r such that $r = p\circ I_r$ for any prior p, H[r,p] is in general the amount of information needed to determine r when given p, or the amount of additional distortion introduced if r is used instead of p [11].

Additional justification for using cross-entropy as a distortion measure in the context of cross-entropy minimization is provided by the following property:

Property B (expected value matching): Let $I(\underset{\sim}{F})$ be the constraints (1) for a fixed set of functions f_k and let $q = p\circ I$ be the result of taking this information into account. Then, for an arbitrary fixed density q*, the cross entropy $H[q^*,q] = H[q^*,p\circ I]$ has its minimum value when the constraints (1) satisfy

$$F_k = F_k^* = \int d\underset{\sim}{x}\; q^*(\underset{\sim}{x}) f_k(\underset{\sim}{x}).$$

This is a generalization of a property of orthogonal polynomials [18, p. 12]

that, in the case of speech analysis, is called the "correlation matching property" [19, Ch. 2]. Property B states that, for a density q of the general form (5), H[q*,q] is smallest when the expectations of q match those of q*. In particular, it follows that q = p•I is not only the density that minimizes H[q,p], as already discussed, but also is the density of the form (5) that minimizes $H[q^{\dagger},q]$! Hence p•I is not only closer to $q^{\dagger}$ than is p -- as shown by the convenient form (11) -- but it is the closest possible density of the form (5).

Another property of cross-entropy minimization that we shall need in Section III is the following:

Property C. Let I_1 and I_2 stand respectively for the constraints

$$\int d\underset{\sim}{x}\, f_i(\underset{\sim}{x}) q_1^{\dagger}(\underset{\sim}{x}) = F_i^{(1)}$$

and

$$\int d\underset{\sim}{x}\, f_i(\underset{\sim}{x}) q_2(\underset{\sim}{x}) = F_i^{(2)} \quad ,$$

which involve the same set of functions f_i, i = 0,...,M. Then

$$(p \circ I_1) \circ I_2 = p \circ I_2 \tag{12}$$

and

$$H[q_2,p] = H[q_2,q_1] + H[q_1,p] + \sum_{r=0}^{M} \beta_r^{(1)} (F_r^{(1)} - F_r^{(2)}) \tag{13}$$

hold, where $q_1 = p \circ I_1$, $q_2 = p \circ I_2$, and the $\beta_r^{(1)}$ are the Lagrangian multipliers associated with $q_1 = p \circ I_1$.

Suppose that $q_1^{\dagger}$ and $q_2^{\dagger}$ are the system probability densities at two different times, and suppose that $q_1^{\dagger}$ or estimates of $q_1^{\dagger}$ are considered to be reasonable prior estimates of $q_2^{\dagger}$. That is, $p \circ I_1$ is considered to be a

reasonable prior estimate of $q_2^\dagger$. Property C states that, when I_2 is determined by expectations of the same functions as I_1, the results of taking I_1 into account are completely wiped out by subsequently taking I_2 into account.

III. CLASSIFICATION METHOD

We now consider the problem outlined in Section I. Let I denote the constraints (1) associated with the feature vector $\underset{\sim}{F}$, and let $\hat{I}_s$ denote analogous constraints associated with each of the pre-defined codewords $\hat{\underset{\sim}{F}}^{(s)}$,

$$\int d\underset{\sim}{x}\ f_i(\underset{\sim}{x}) q_s^\dagger(\underset{\sim}{x}) = \hat{F}_i^{(s)} \quad . \tag{14}$$

Suppose that p is an estimate of $q^\dagger$ that is available prior to learning $\underset{\sim}{F}$. Then the best posterior estimate of $q^\dagger$ is

$$q = p \circ I \quad , \tag{15}$$

in the sense discussed in Section II. Now, let $\hat{q}_s = p \circ \hat{I}_s$ be the minimum cross-entropy estimates of $q^\dagger$ that would apply if the current feature vector $\underset{\sim}{F}$ were equal to the codeword $\hat{\underset{\sim}{F}}^{(s)}$. As discussed in Section II, $H[q,\hat{q}_s]$ is the amount of information-theoretic distortion introduced if q is represented by $\hat{q}_s$. It is therefore reasonable to define the distortion measure between $\underset{\sim}{F}$ and $\hat{\underset{\sim}{F}}^{(s)}$ as

$$D(\underset{\sim}{F}, \hat{\underset{\sim}{F}}^{(s)}) = H[q,\hat{q}_s], \tag{16}$$

where $q = p \circ I$ and $\hat{q}_s = p \circ \hat{I}_s$. The nearest neighbor classification rule (2) then becomes: find t such that

$$H[q,\hat{q}_t] = \min_{s\in\Lambda} H[q,\hat{q}_s] \quad , \tag{17}$$

if the minimum exists (e.g., if Λ is finite). Now, from (11), we have

$$H[q^\dagger,\hat{q}_s] = H[q^\dagger,q^*] + H[q^*,\hat{q}_s] \quad , \tag{18}$$

where $q^* = \hat{q}_s \circ I$ is the estimate of $q^\dagger$ that results if $\hat{q}_s$ instead of p is used in (15) as the prior estimate of $q^\dagger$. But

$$q^* = \hat{q}_s \circ I = (p\circ\hat{I}_s)\circ I = p\circ I = q \tag{19}$$

follows from (12), so that (18) becomes

$$H[q^\dagger,\hat{q}_s] = H[q^\dagger,q] + H[q,\hat{q}_s] \quad . \tag{20}$$

$H[q^\dagger,\hat{q}_s]$ is the amount of information needed to determine $q^\dagger$ given the codeword $\hat{q}_s$, or the amount of distortion introduced when $q^\dagger$ is represented by $\hat{q}_s$. Equation (20) states that the total distortion $H[q^\dagger,\hat{q}_s]$ is the sum of the distortion introduced when $q^\dagger$ is represented by the best posterior estimate $q = p\circ I$ plus the distortion introduced when q is represented by the codeword $\hat{q}_s$. Since q minimizes $H[q^\dagger,q]$ in the sense defined in Property B, (20) shows that the classification rule (17) is optimal in the sense of minimizing the total distortion $H[q^\dagger,\hat{q}_s]$. The rule (17) is equivalent to the minimum discrimination classification method of Kullback [6, p. 83] since $q = \hat{q}_s \circ I$ by (19), which shows that the Kullback method is optimal in a sense that has not been appreciated previously. Notice that when q is in the codeword set -- $q = \hat{q}_s$ for some $s\in\Lambda$ -- the rule (17) just selects q, the best posterior estimate of $q^\dagger$.

Minimizing $H[q,\hat{q}_t]$ identifies the associated codeword $\hat{\underline{x}}^{(t)}$. Now, the quantity $H[q,\hat{q}_t] = H[\hat{q}_t\circ I,\hat{q}_t]$ is the amount of information provided by I

that was not already inherent in $\hat{q}_t$. We can therefore restate the solution in terms of the problem as originally posed -- choosing the codeword $\hat{\underset{\sim}{F}}^{(t)}$ that "best represents" the feature vector $\underset{\sim}{F}$ -- in the following way: Choose the codeword $\hat{\underset{\sim}{F}}^{(t)}$ such that the feature vector $\underset{\sim}{F}$ provides the least additional information beyond what $\hat{\underset{\sim}{F}}^{(t)}$ provides.

We now consider the computational requirements of the classification method. At this point we specialize to the case in which there is a discrete set of n codewords $\hat{\underset{\sim}{F}}^{(j)}$. Given an input feature vector $\underset{\sim}{F}$ of M+1 measurements F_i, the classification procedure may be summarized as follows:

a) compute $q = p \circ I$, where I represents the constraints (1);

b) compute $H[q,\hat{q}_j]$ $(j = 1,\ldots,n)$, where $\hat{q}_j = p \circ \hat{I}_j$ and $\hat{I}_j$ represents the constraints (14), and find a value j such that $H[q,\hat{q}_j] \leq H[q,\hat{q}_i]$, for $i \neq j$.

Now, owing to property C, it turns out that the first step is unnecessary: this "two-step" procedure reduces to a single step. From (13), it follows that

$$H[q,\hat{q}_j] = H[q,p] - H[\hat{q}_j,p] - \sum_{k=0}^{M} \hat{\beta}_k^{(j)}(\hat{F}_k^{(j)} - F_k)$$

or

$$H[q,\hat{q}_j] = H[q,p] + \hat{\lambda}^{(j)} + \sum_{k=0}^{M} \hat{\beta}_k^{(j)} F_k \qquad (21)$$

holds, where we have substituted for $H[\hat{q}_j,p]$ by means of (6). In (21) the F_k are components of the the input feature vector $\underset{\sim}{F}$ and the $\hat{\lambda}^{(j)}$ and $\hat{\beta}_k^{(j)}$ are Lagrangian multipliers associated with $\hat{q}_j = p \circ \hat{I}_j$. Since the $\hat{I}_j$ are known ahead of time, these multipliers can be computed ahead of time [11, Appendix A],[16]. Now the quantity $H[q,p]$ is a constant for any feature vector $\underset{\sim}{F}$, so that the closeat codeword $\hat{\underset{\sim}{F}}^{(j)}$ can be determined by finding the

smallest of the quantities

$$\Delta_j = \hat{\lambda}^{(j)} + \sum_{k=0}^{M} \hat{\beta}_k^{(j)} F_k \quad , \tag{22}$$

which does not involve having to compute $q = p \cdot I$. Computing each Δ_j requires M+1 multiplications and additions involving M+2 pre-stored multipliers and the M+1 elements of the input vector $\underset{\sim}{F}$. If there are n possible codewords, the total requirement is n(M+1) multiplications and additions, storage for n(M+2) Lagrangian multipliers, and approximately n comparisons (to find the smallest Δ_j). One can also trade about n(M+1)/2 of the multiplications for additions [20]. Since the Δ_j can be computed independently, concurrent computation is possible.

These results are a generalization of the method of speech coding by vector quantization [3],[2], which exploits a special case of (20) that was found to hold for a speech spectral distortion measure due to Itakura and Saito [21],[22]. Under suitable assumptions, the Itakura-Saito distortion measure can be shown to be a special case of assymptotic cross-entropy rate [2],[22]. In Section V, we show how speech coding by vector quantization follows as a special case from (22).

IV. COMPUTATION OF CLUSTER CENTROIDS

Suppose that a cluster of measurement vectors $\underset{\sim}{F}^{(i)}$, i = 1,...,N, is to be represented in the classification procedure of Section III by a single, "centroid" codeword $\hat{\underset{\sim}{F}}$. For example, the $\underset{\sim}{F}^{(i)}$ might result from measurements on N members of the class to be represented by $\hat{\underset{\sim}{F}}$. How should one determine $\hat{\underset{\sim}{F}}$ from the $\underset{\sim}{F}^{(i)}$?

The selection of centroids is a key facet of cluster analysis techniques such as the k-means technique [5] or the ISODATA technique [23], and it is

also important in the design of vector quantizers [24]. When the distortion measure is the Euclidean distance, centroids are simply Euclidean centers of gravity. For more general distortion measures as in (16), a natural generalization of the Euclidean centroid [24] of a collection $\{\underset{\sim}{F}^{(i)};\ i=1...N\}$ is the vector $\hat{\underset{\sim}{F}}$ minimizing the average distortion,

$$D_c(\hat{\underset{\sim}{F}}) = \frac{1}{N}\sum_{i=1}^{N} D(\underset{\sim}{F}^{(i)},\hat{\underset{\sim}{F}}) = \frac{1}{N}\sum_{i=1}^{N} H[q_i,\hat{q}] \quad , \tag{23}$$

where $q_i = p \circ I_i$ and $\hat{q} = p \circ \hat{I}$. Here, I_i and $\hat{I}$ stand for expected value constraints of the form (1) corresponding respectively to $\underset{\sim}{F}^{(i)}$ and $\hat{\underset{\sim}{F}}$. Perhaps surprisingly, the centroid for this apparantly complicated non-Euclidean distortion measure can be readily evaluated because of the special properties of Section II. In fact, we show below that the minimum of $D_c(\hat{\underset{\sim}{F}})$ is achieved simply by the components of $\hat{\underset{\sim}{F}}$ each being the arithmetic mean of the components of the $\underset{\sim}{F}^{(j)}$.

Since I_i and $\hat{I}$ involve the same constraint functions f_j, Property C applies. Eq. (13) yields

$$H[q_i,\hat{q}] = H[q_i,p] - H[\hat{q},p] - \sum_{r=0}^{M}\hat{\beta}_r(\hat{F}_r - F_r^{(i)}) \quad ,$$

where the $\hat{\beta}_r$ are Lagrangian multipliers associated with $q = p \circ \hat{I}$. It follows that (23) becomes

$$D_c(\hat{\underset{\sim}{F}}) = \frac{1}{N}\sum_{i=1}^{N} H[q_i,p] - H[\hat{q},p] - \sum_{r=0}^{M}\hat{\beta}_r(\hat{F}_r - \overline{F}_r) \ , \tag{24}$$

where the $\overline{F}_r$ are components of the mean constraints

$$\overline{\underset{\sim}{F}} = \frac{1}{N}\sum_{i=1}^{N} \underset{\sim}{F}^{(i)} \quad . \tag{25}$$

Now, let $\overline{q} = p \circ \overline{I}$, where $\overline{I}$ represents the mean constraints $\overline{\underset{\sim}{F}}$. Then (13) yields

$$H[\bar{q},\hat{q}] = H[\bar{q},p] - H[\hat{q},p] - \sum_{r=0}^{M} \hat{\beta}_r(\hat{F}_r - \bar{F}_r)$$

By combining this with (24), we obtain

$$D_c(\hat{\underset{\sim}{F}}) = H[\bar{q},\hat{q}] - H[\bar{q},p] + \frac{1}{N}\sum_{i=1}^{N} H[q_i,p] \tag{26}$$

Since $D_c(\hat{\underset{\sim}{F}})$ depends on $\underset{\sim}{F}$ only through the first term, minimizing $D_c(\hat{\underset{\sim}{F}})$ is equivalent to minimizing $H[\bar{q},\hat{q}]$. This minimum occurs when $\bar{q} = \hat{q}$ (see (10), which in turn means that the optimal centroid $\hat{\underset{\sim}{F}}$ is $\hat{\underset{\sim}{F}} = \bar{\underset{\sim}{F}}$, where $\bar{\underset{\sim}{F}}$ is given by (25). Hence the components of the cluster centroid $\underset{\sim}{F}$ are just the arithmetic means of the components of the cluster elements $\underset{\sim}{F}^{(j)}$.

V. EXAMPLE --- SPEECH CODING BY VECTOR QUANTIZATION

Speech coding by vector quantization is a recently developed narrow-bandwidth speech coding technique based on Linear Prediction Coding (LPC) [3],[2]. Based on estimates of the sample autocorrelation function that are measured in each frame, the speech in each frame is coded in terms of the identity of a prestored set of LPC parameters called a codeword. The LPC paramaters used are the inverse filter gain σ^2 and sample coefficients a_i, $i = 0,\ldots,M$, with

$$a_0 = 1 \quad . \tag{27}$$

These parameters characterize a filter that is used in synthesizing the speech after decoding. The nearest-neighbor distortion rule used in coding the speech selects in each frame the codeword that has the smallest Itakura-Saito distortion [21],[22] with respect to the current frame of speech. In particular ([2],[3]), one finds the codeword with parameters that minimize the

expression

$$\frac{1}{\sigma^2}\left\{r_x(0)r_a(0) + 2\sum_{s=1}^{M} r_x(s)r_a(s)\right\} + \log(\sigma^2) \quad , \tag{28}$$

where

$$r_a(s) = \begin{cases} \sum_{i=0}^{M-s} a_i a_{i+s} & , \quad s \leq M \\ 0 & , \text{ otherwise} \end{cases} \quad , \tag{29}$$

and where the $r_x(s)$ are measurements that estimate the autocorrelation function of the speech in the current frame for lags $s = 0,1,\ldots,M$. For convenience, we omit indexing σ and the a_i with the codebook parameter j over which (28) is minimized.

Gray, et al. [2] have shown that the minimization expression (28) can be derived by means of cross-entropy minimization. Here, we shall show that (28) is a special case of the general expression (22). In doing so, we shall use some results from [2]. In [2], the derivation is conducted in terms of codebook probability densities $\hat{q}(\underset{\sim}{y})$, where $\underset{\sim}{y} = y_0,\ldots,y_{k-1}$ is a vector of k time domain signal samples. In particular, each codebook entry corresponds to an autogregressive model of the form

$$y_i = u_i \qquad (i = -M,\ldots-1)$$

$$y_n = -\sum_{j=1}^{M} a_j y_{n-j} + \sigma e_n \quad ,$$

where the u_i are the initial conditions for the filter, where the y_n are time-domain signal samples, and where e_n is a zero-mean, unit-variance sequence of independent Gaussian random variables. The vector $\underset{\sim}{y}$ can be expressed in the form $\underset{\sim}{y} = \underset{\approx}{A}^{-1}(\sigma\underset{\sim}{e} + \underset{\sim}{y})$, where $\underset{\approx}{A}$ is a banded,

triangular matrix whose components are the inverse filter sample coefficients,

$$(\underset{\sim}{A})_{ij} = \begin{cases} a_{i-j} , & 0 \leq i-j \leq M \\ 0 , & \text{otherwise} \end{cases} , \tag{30}$$

and where

$$(\underset{\sim}{v})_n = \begin{cases} -\sum_{j=n-M}^{-1} a_{n-j} u_j & n < M \\ 0, & n \geq M \end{cases} \tag{31}$$

Each codebook density $\hat{q}(\underset{\sim}{y})$ is Gaussian,

$$\hat{q}(\underset{\sim}{y}) = (2\pi)^{k/2} (\det \underset{\sim}{R}_2)^{-1/2} \exp[-(\underset{\sim}{y} - \underset{\sim}{m})^t \underset{\sim}{R}_2^{-1} (\underset{\sim}{y} - \underset{\sim}{m})/2] , \tag{32}$$

with mean

$$\underset{\sim}{m} = E(\underset{\sim}{y}) = \underset{\sim}{A}^{-1} \underset{\sim}{v} , \tag{33}$$

and covariance

$$\underset{\sim}{R}_2 = E((\underset{\sim}{y}-E(\underset{\sim}{y}))(\underset{\sim}{y}-E(\underset{\sim}{y}))^t)$$

$$= \sigma^2 (\underset{\sim}{A}^t \underset{\sim}{A})^{-1} , \tag{34}$$

where E stands for expected value and where t indicates a transpose operation. For convenience, we omit the voicing parameter that is part of the analysis in [2]. Our results, however, are unaffected by this omission.

In [2], the expression (28) is obtained by using constraints

$$E(y_i y_j) = r_x(|i-j|) , \quad |i-j| \leq M \tag{35}$$

$$E(\underset{\sim}{y}) = 0 \tag{36}$$

for each speech frame, applying Kullback's classification procedure [6, p. 83] to select a codebook entry $\hat{q}$, and taking the $k \to \infty$ limit of the per-symbol cross-entropy so that the classification is based on the stable, non-transient behavior of the autoregressive models.

The codebook densities (32) were derived in [2] directly from arguments concerning the speech reproduction model class. For our purpose here -- applying the results of Section III -- we need to express the codebook densities as minimum cross-entropy densities $\hat{q} = p \circ \hat{I}$, for some fixed prior p and codebook-dependent constraints $\hat{I}$. This is accomplished ([2],[6],[10],[11]) by the prior

$$p(\underset{\sim}{y}) = (2\pi)^{-k/2} \exp[-\underset{\sim}{y}^t \underset{\sim}{I} \underset{\sim}{y}/2] \quad , \tag{37}$$

where $\underset{\sim}{I}$ is the identity matrix, and by the constraints

$$E(y_i y_j) = (\underset{\sim}{R}_2 + \underset{\sim}{m}^t \underset{\sim}{m})_{ij} \quad , \qquad |i-j| \leq M \tag{38}$$

$$E(\underset{\sim}{y}) = \underset{\sim}{m} \quad . \tag{39}$$

We can now make the connection with the results of Section III. There we showed that the best codeword is determined by the minimum of the quantities (22),

$$\Delta = \hat{\lambda} + \sum_k \hat{\beta}_k F_k \quad , \tag{40}$$

which we have rewritten here without the codebook index j. In terms of the foregoing, the measurements F_k are given by the right hand sides of (35)-(36), the $\hat{\beta}_k$ are the Lagrangian multipliers in $\hat{q} = p \circ \hat{I}$ that correspond to the constraints (38)-(39), and $\hat{\lambda}$ is the normalization multiplier. The Lagrangian multipliers can be identified in (32) by noting that $\hat{q}$ must have the general form (5). After factoring out the prior (37), it is easy to see that $\hat{\lambda}$ is given by

$$\exp[-\lambda] = (\det \underset{\sim}{R}_2)^{-1/2} \exp[-\underset{\sim}{m}^t \underset{\sim}{R}_2^{-1} \underset{\sim}{m}/2] \tag{41}$$

and that the terms $y_i y_j$ in the exponential in (32) have the factors $(1/2)(\underset{\sim}{R}_2^{-1} - \underset{\sim}{I})_{ij}$, which are therefore the Lagrangian multipliers corresponding to the constraints (38). We do not need the multipliers corresponding to (39) since, owing to (36), they do not contribute to the sum

$\sum_k \hat{\beta}_k F_k$ in (40). Eq. (40) therefore becomes

$$\Delta = \hat{\lambda} + \frac{1}{2} \sum_{|i-j| \le M} r_x(|i-j|)(\underset{\sim}{R}_2^{-1} - \underset{\sim}{I})_{ij}$$

$$= \hat{\lambda} + \frac{1}{2} \sum_{|i-j| \le M} r_x(|i-j|)(\underset{\sim}{R}_2^{-1})_{ij} - \frac{k}{2} r_x(0) \quad . \tag{42}$$

The last term cannot affect the minimization since it doesn't depend on the codebook entry; we therefore drop it. From (39), we have

$$\hat{\lambda} = \frac{1}{2} \underset{\sim}{m}^t \underset{\sim}{R}_2^{-1} \underset{\sim}{m} + \frac{1}{2} \log \det \underset{\sim}{R}_2$$

$$= \frac{1}{2} \underset{\sim}{m}^t \underset{\sim}{R}_2^{-1} \underset{\sim}{m} + \frac{1}{2} \log \det \sigma^2 (\underset{\sim}{A}^t \underset{\sim}{A})^{-1}$$

$$= \frac{1}{2} \underset{\sim}{m}^t \underset{\sim}{R}_2^{-1} \underset{\sim}{m} + \frac{k}{2} \log \sigma^2 + \frac{1}{2} \log \det \underset{\sim}{A} + \frac{1}{2} \log \det \underset{\sim}{A}^t$$

$$= \frac{1}{2} \underset{\sim}{m}^t \underset{\sim}{R}_2^{-1} \underset{\sim}{m} + \frac{k}{2} \log \sigma^2$$

where we have used (34) and, in the last step, the fact that $\underset{\sim}{A}$ is triangular with $A_{ii} = a_0 = 1$ from (30) and (27). Since minimizing Δ in (42) is equivalent to minimizing $2\Delta/k$, it follows that the best codeword can be found by minimizing

$$\Gamma = \frac{1}{k} \underset{\sim}{m}^t \underset{\sim}{R}_2^{-1} \underset{\sim}{m} + \log(\sigma^2) + \frac{1}{k} \sum_{|i-j| \le M} r_x(|i-j|)(\underset{\sim}{R}_2^{-1})_{ij} \tag{43}$$

Now, the first term in (43) evaluates to $(\underset{\sim}{v}^t \underset{\sim}{v})/k\sigma^2$ by means of (33)-(34). Since $\underset{\sim}{v}$ has a fixed number of M+1 non-zero terms (see (31)), it follows that this term goes to zero as $k \to \infty$.

Expanding (34) by means of (30) leads to

$$(\underset{\sim}{R}_2^{-1})_{ij} = \frac{1}{\sigma^2} \sum_{n=\max(i,j)}^{\min(i+M, j+M, k-1)} a_{n-i} a_{n-j} \tag{44}$$

Since $\underset{\sim}{R}_2^{-1}$ is symmetric, it suffices to consider the case $i > j$, for which

(44) becomes

$$(\underset{\sim}{R}_2^{-1})_{ij} = \frac{1}{\sigma^2} \sum_{n=i}^{\min(M+j,k-1)} a_{n-i} a_{n-j}$$

$$= \frac{1}{\sigma^2} \sum_{s=0}^{\min(M-|i-j|,k-1-i)} a_s a_{s+|i-j|} \quad .$$

Provided that

$$\min(i,j) < k-1-M \tag{45}$$

holds,

$$(\underset{\sim}{R}_2^{-1})_{ij} = \frac{1}{\sigma^2} r_a(|i-j|) \quad , \quad |i-j| \leq M \quad ,$$

follows from (29). Equation (43) is therefore equivalent to

$$\Gamma = \log(\sigma^2) + \frac{1}{k} \sum_{|i-j| \leq M} r_x(|i-j|) r_a(|i-j|)$$

$$= \log(\sigma^2) + \frac{1}{\sigma^2} \left\{ r_x(0) r_a(0) + 2 \sum_{s=1}^{M} r_x(s) r_a(s) \right\} \quad , \tag{46}$$

which is the same as (22). In deriving the last term of (46), we have ignored corrections necessary for proper evaluation of the sum at the matrix boundaries. However, these corrections involve only $(M+1)^2$ terms (see (45)) and therefore become negligible as $k \to \infty$.

VI. GENERAL DISCUSSION

The special properties of cross-entropy that hold for minimum cross-entropy densities [11] result in a pattern classification method with several advantages: It is optimal in a well-defined, information-theoretic sense; it is computationally attractive; and it includes a self-consistent,

simple method of computing the set of cluster centroids in terms of which the classification is made. A special case of this method (speech coding by vector quantization) has already proved to be successful. It therefore seems likely that the method can be used successfully in a variety of other applications.

VII. ACKNOWLEDGMENT

We thank Rodney Johnson for helpful discussions and for his review of an earlier manuscript.

References

1. J.E. Shore, "Minimum Cross-Entropy Spectral Analysis," NRL Memorandum Report 3921, Naval Research Laboratory, Washington, D.C. 20375, January, 1979.

2. R.M. Gray, A.H. Gray, Jr., G. Rebolledo, and J.E. Shore, "Rate-Distortion Speech Coding With a Minimum Discrimination Information Distortion Measure," submitted to IEEE Trans. Information Theory.

3. A. Buzo, A.H. Gray, Jr., R.M. Gray, and J.D. Markel, "Speech Coding Based Upon Vector Quantization," submitted to IEEE Trans. Acoustics, Speech, and Signal Processing.

4. J.A. Hartigan, Clustering Algorithms, New York, John Wiley, 1975.

5. J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symposium on Math., Stat., and Prob., vol. I, Univ. Ca. Press, 1967, pp. 281-286.

6. S. Kullback, Information Theory and Statistics, New York, Dover, 1969.

7. M.S. Pinsker, Information and Information Stability of Random Variables and Processes, San Francisco, Holden-Day, 1964.

8. R.W. Johnson, "Axiomatic Characterization of the Directed Divergences and Their Linear Combinations," IEEE Trans. Information Theory IT-25, Nov. 1979, pp. 709-716.

9. J.E. Shore and R.W. Johnson, "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," IEEE Trans. Information Theory IT-26, Jan. 1980.

10. I. Csiszar, "I-Divergence Geometry of Probability Distributions and Minimization Problems," Ann. Prob. 3, No. 1, 1975, pp. 146-58.

11. J.E. Shore and R.W. Johnson, "Properties of Cross-Entropy Minimization," NRL Memorandum Report 4189, Naval Research Laboratory, Washington, D.C. 20375, March 1980, also submitted to IEEE Trans. Information Theory.

12. I.J.Good, "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," Annals Math. Stat. 34, 1963, pp. 911-934.

13. I.J. Good, Probability and the Weighing of Evidence, Charles Griffen, London, 1950.

14. W.M. Elsasser, "On Quantum Measurements and the Role of the Uncertainty Relations in Statistical Mechanics," Phys. Rev. 52, (Nov. 1937), pp. 987-999.

15. E.T. Jaynes, "Information Theory and Statistical Mechanics I," Phys. Rev. 108, 1957, pp. 171-190.

16. R.W. Johnson, "Determining Probability Distributions by Maximum Entropy and Minimum Cross-Entropy," Proceedings APL79, (ACM 0-89791-005), May 1979, pp. 24-29.

17. N. Jardeen and R. Sibson, Mathematical Taxonomy, New York, John Wiley, 1971.

18. L. Geronimus, Orthogonal Polynomials, New York, Consultants Bureau, 1961.

19. J.D. Markel and A.H. Gray, Jr., Linear Prediction of Speech, New York, Springer-Verlag, 1976

20. S. Winograd, "A New Algorithm for Inner Products," IEEE Trans. Computers C-17, 1968, pp. 693-694.

21. F. Itakura and S. Saito, "Analysis Synthesis Telephone Based Upon the Maximum Lilkihood Method", Reports of the 6th Int. Cong. Acoustics, Y. Yonasi, ed., Tokyo, 1968.

22. R.M. Gray, A. Buzo, A.H. Gray, Jr., and Y. Matsuyama, "Distortion Measures for Speech Processing," IEEE Trans. Acoustics, Speech, and Signal Processing, to appear.

23. G.H. Ball and D.J. Hall, "Isodata - An Iterative Method of Multivariate Analysis and Pattern Classification," Proc. IFIPS Congress 1965.

24. Y. Linde, A. Buzo, and R.M. Gray, "An Algorithm for Vector Quantizer Design," IEEE Trans. Comm., to appear.

END

DATE
FILMED

8-80

DTIC